International Electronic Journal of
Mathematics Education
Volume 2, Number 3, October 2007 www.iejme.com

EXPLORING CONNECTIONS BETWEEN SAMPLING DISTRIBUTIONS AND STATISTICAL INFERENCE: AN ANALYSIS OF STUDENTS' ENGAGEMENT AND THINKING IN THE CONTEXT OF INSTRUCTION INVOLVING REPEATED SAMPLING

Luis Saldanha

Patrick Thompson

ABSTRACT. Construing a collection of values of a sample statistic as a distribution is central to developing a coherent understanding of statistical inference. This paper discusses key developments that unfolded over three consecutive lessons in a classroom teaching experiment designed to support a group of high school students in developing such a construal. Instruction began by engaging students in activities that focused their attention on the variability among values of a common sample statistic. There occurred a critical shift in students' attention and discourse away from individual values of the statistic and toward a collection of such values as a basis for inferring the value of a population parameter. This was followed by their comparisons of such collections and by the emergence and application of a rule for deciding whether two such collections were similar. In the repeated application of their decision rule students structured these collections as distributions. We characterize aspects of these developments in relation to students' classroom engagement, and we explore evidence in students' written work that points to how instruction shaped their conceptions.

KEYWORDS. Sampling, Variability, Sample Statistic, Population Parameter, Statistical Inference, Sampling Distributions, Teaching Experiment, Conceptual Operations.

INTRODUCTION

Statistical inference is one of the most sophisticated and important schemes of ideas in introductory statistics. While data analysis techniques focus on the information content and structure of specific collections of data, it is statistical inference that situates data with respect to a population from which it might be drawn, allowing us to examine the extent to which information extracted from the data in hand can be generalized to the population. The importance to everyday citizenship of understanding statistical inference is clear. Citizens are confronted frequently with published reports of opinion surveys, justifications and implications of policy,

ISSN: 1306-3030

and reports of drug trials and experimental medical treatments. Citizens are also confronted with conflicting reports, thus ideas of sampling and statistical inference are important for understanding the degree to which data-based claims are warranted and for understanding that conflicting claims are not necessarily a sign of confusion or duplicity.

With regard to the teaching of statistical inference, the National Council of Teachers of Mathematics (NCTM, 2000) includes the following recommendations for grades 9 to 12. Students should:

- Use simulations to explore the variability of sample statistics from a known population and to construct sampling distributions;
- Understand how sample statistics reflect the values of population parameters and use sampling distributions as the basis for informal inference. (NCTM, 2000, p. 324) .

Though the NCTM's recommendations seem straightforward, the presence of the term "understand" in their statement makes them problematic. It is unclear in these recommendations what it means to understand how sample statistics reflect the values of population parameters. Nor is it clear what it means to understand the use of sampling distributions as the basis for informal inference. It is not even clear what it means to understand sampling distributions. This lack of clarity and specificity is problematic because without models of what it might mean to understand these important ideas, practitioners' abilities to identify and assess students' understandings of them are severely compromised, as are researchers' abilities to design effective curricula for supporting students' understanding of them.

In this paper we discuss insights into what it can mean to understand some of these ideas. Our insights are drawn from data generated in a classroom instructional experiment designed to support high school students in exploring and coming to construe important connections between the ideas of statistical inference and sampling distributions. The paper begins by providing the background to our study in the form of a directed analysis of a relevant body of research on statistics learning and instruction. We then describe the study, situating its purpose within a larger study and its goals, providing information about our participants, and elaborating our method of inquiry and the instructional context. Our results are then presented in two parts: the first part frames the class as a unit of analysis and gives a developmental account of the unfolding of instruction in tandem with students' engagement; the second part discusses students' written work, drawing inferences about individual students' underlying conceptions and relating those back to the account in the first part. We conclude with a summary of our findings, highlighting their significance and drawing out a limitation of the study that points to future research.

PRIOR RELEVANT RESEARCH

Singular versus Distributional Perspectives on Sampling

There is substantial evidence in the research literature that students, when asked to make judgments about outcomes of random sampling, tend to focus on individual samples and statistical summaries of them instead of how collections of sample statistics are distributed. For instance, Kahneman and Tversky (1972) produced empirical evidence to support their hypothesis that people often base judgments of the probability that a sample will occur on the degree to which they think the sample "(i) is similar in essential characteristics to its parent population; and (ii) reflects the salient features of the process by which it is generated" (ibid., p. 430). In later research, Kahneman and Tversky (1982) conjectured that people tend to take a singular rather than a distributional perspective when making judgments under uncertainty. A singular perspective is characterized by a focus on the causal system that produced the particular outcome and by an assessment of likelihood based on "the propensities of the particular case at hand". In contrast, a distributional perspective relates the case at hand to a sampling schema and views an individual case as "an instance of a class of similar cases, for which relative frequencies of outcomes are known or can be estimated" (ibid, p. 518).

Konold (1989) found strong empirical support for Kahneman and Tversky's (1982) conjecture. He presented compelling evidence that people, when asked questions that are ostensibly about probability, interpret such questions as asking to predict with certainty the outcome of an individual trial of an experiment. The participants in Konold's study often based their predictions of random sampling outcomes on causal explanations instead of information obtained from repeating an experiment. Konold (ibid.) referred to these combined orientations as the outcome approach. Moreover, he noted that his participants adhered strongly to this approach, even in the face of evidence designed to impel them to change their perspective. Decades earlier, Piaget & Inhelder (1951) documented similar robust orientations among young children who participated in experiments designed to query their intuitions and conceptions related to chance and likelihood.

Sedlmeier and Gigerenzer (1997) conducted an extensive analysis of decades of research on the effects of sample size on statistical prediction. They concluded that participants across a diverse spectrum of studies who incorrectly answered tasks involving a distribution of sample statistics probably interpreted task situations and questions as being about individual samples.

Research on Understanding Sampling Distributions

Some notable studies have addressed students' understanding of ideas more directly related to sampling distributions within instructional settings (Well, Pollatsek, & Boyce, 1990; delMas, Garfield, & Chance, 1999; Sedlmeier, 1999, Saldanha & Thompson, 2002; Thompson,

Saldanha & Liu, 2004). delMas et al. (1999) conducted a three-phase design experiment that used a computer environment in which students could simulate drawing large numbers of samples from populations having given distributions, and in which they could vary the size of samples and view the resulting distribution of the values of a statistic calculated from the samples. The researchers found in the first phase that students often anticipated that larger sample size should produce a distribution of sample statistics that resembles the parent population. Even with two subsequent rounds of refinement of both the computer tool and learning activities with it, there were students who still held that same expectation tenaciously. One possible explanation is that students held the intuition that a mean is a data point within the population, and therefore gathering more means should result in a distribution that strongly resembles the population.

Sedlmeier (1999, pp. 128-139) explored the effects of a four-phase training program involving a micro-world that students used to simulate sampling from a virtual urn containing red and blue balls. The initial phase had students observe demonstrations of different-sized samples drawn repeatedly from the urn and their respective "stack plot" of sample proportions. The final phase had students model a variant of Kahneman and Tversky's (1972) Maternity Ward problem, in the accompaniment of textual explanations generated by the micro world about issues to which they should attend when considering the effect of different sample sizes on the distribution of proportions. Sedlmeier reported that students' performance on target tasks isomorphic to this problem improved over the four phases.

Well, Pollatsek and Boyce (1990) studied college students' understanding of the arithmetic mean's variability. They used a computer display of a population's distribution alongside displays of distributions of sample statistics calculated from a large number of small samples and a large number of large samples. Though the interviewer engaged in a teaching interview (pointing out similarities and differences among the distributions), many students did not realize that the distribution of means for large samples is less variable than that for small samples.

Saldanha and Thompson (2002) and Thompson, Saldanha and Liu (2004) conducted a teaching experiment in which they engaged a class of high school students in making informal inferences on the basis of distributions of a sample statistic. Instruction entailed having students employ, describe the operation of, and explain the results of computer simulations of taking large numbers of samples from various populations with known parameters. The experiment concluded by having students examine simulation results and frequency histograms of them systematically, with the aim that they explore how distributions of sample proportions are affected by underlying population proportions and by sample size. The authors reported (Saldanha & Thompson, 2002) that many students experienced significant and robust difficulties in conceptualizing sampling distributions as having a multi-tiered and hierarchical structure

involving the selection of individual items to form a sample, on one level, and then repeating this process and computing the value of a sample statistic to form a collection of values (each one denoting the composition of a single sample), on another level. In particular, students would persistently and easily loose track of the repeated sampling process; their thinking would often unwittingly shift from focusing on a number of items in a sample to a number of samples selected. Their control of the coordination between the various levels of imagery was unstable; from one moment to the next their image of a number of samples (of items) seemed to easily dissolve into an image of a total number of items in all the samples. This muddling of the different levels of the repeated sampling process, in turn, obstructed their abilities to imagine how sample proportions might distribute themselves around the underlying population proportion.

The Role of Variability

The ubiquity of variability among outcomes generated by random processes is a central idea in statistics (Cobb & Moore, 1997). Despite its centrality, students' understanding of variability and its potential role as an organizing idea in statistics instruction have received relatively little research attention (Shaughnessy, Watson, Moritz, & Reading, 1999). Rubin, Bruce and Tenney (1991) elaborated a conceptual analysis in which they proposed that the integration of two seemingly contrasting ideas underlies a coherent understanding of sampling and inference: 1) *sampling representativeness* - the expectation that a sample taken from a population will often have characteristics similar to that population's, and 2) *sampling variability* - the expectation that different samples selected from a common population will differ among each other and from the sampled population. Rubin et al.'s (ibid.) investigation of statistically untrained high school students' reasoning on sampling and inference tasks showed that students did not integrate these two ideas to reason about distributions of sample outcomes. Instead, one or the other expectation seemed more salient in students' minds, depending on the task. Rubin et al.'s elaboration of sampling variability points implicitly at the centrality of the idea of repeated sampling, but the idea is not at the foreground of their conceptual analysis. Similarly, repeated sampling was not part of the student tasks employed in their study.

Schwartz, Goldman, Vye and Barron (1998) and Watson and Moritz (2000) both characterized sampling as a method of indirectly obtaining information about a larger population by directly obtaining information from only a relatively small and representative subset of the population. Neither characterization, however, entailed images of the repeatability of the sampling process nor of the variability that we can expect among sampling outcomes.

Summary

The relationship of the above literature to issues of understanding statistical inference is that gathering a sample and calculating a statistic from it can be viewed as a *stochastic event*. One conceives of an event as being stochastic when one understands it to be generated by a repeatable random process which is expected to produce variable outcomes of the event. In sampling contexts, if one views collecting a sample as gathering information about an attribute of the underlying population then the outcome of each sample is seen to estimate the population attribute by quantifying that of the sample instead. Combining these two ideas one can anticipate that the values of a sample statistic will vary somewhat under repeated sampling and that aggregates of such values will naturally be internally "diverse". That people often conceive (what we see as) stochastic events non-stochastically has important implications for how they draw inferences and how they understand instruction aimed at developing a normative understanding of statistical inference. For instance, we do not see how the normative practice of drawing an inference from an individual sample to a population can be understood deeply without reconciling the ideas of sample-to-sample variability and relative frequency patterns that emerge in collections of values of a sample statistic—ideas for which a stochastic conception of sampling is foundational.

Following this last assertion and the perspective of a stochastic conception of sampling, Saldanha and Thompson (2002) have argued that singular interpretations of likelihood (Kahneman & Tversky, 1972, 1982; Konold, 1989) and conceptions of sampling that do not foreground ideas of variability and the repeatability of the sampling process are problematic for learning statistical inference because they disable one from considering the relative unusualness of a sampling process' outcome. Drawing on the results discussed here and on data from a teaching experiment, Saldanha and Thompson characterized a conception of sampling that entails images of the repeatable sampling process, the bounded variability among values of a sampling statistic, and the expected fuzzy similarity between sample and population. Our characterization links these images schematically to form an organized system of ideas that, we claim, supports building deep connections between sampling and inference. We thus propose this scheme-based conception of sampling as a powerful and enabling instructional endpoint.

THE STUDY

Purpose

Our report is part of a larger study (Saldanha, 2004) that investigated students' abilities to conceive the ideas of variability, samples, and sampling distributions as an interrelated scheme. The report focuses on events that transpired over three consecutive classroom lessons in the initial phase of a classroom teaching experiment conducted within an intact introductory statistics course at the secondary school level. Our instructional intent in this initial phase was to

lay the conceptual groundwork for students to develop a coherent understanding of the concept of margin of error, which we assert is supported by students having an image of frequency patterns that emerge as values of a sample statistics accumulate and disperse themselves along a range of possible values under repeated sampling from a common population. We treated margin of error explicitly only near the end of the experiment, and the concept per se is beyond the scope of this report. However, it is important to communicate that our aim in the initial phase was to have students understand the logic of making inferences to a population on the basis of how a collection of values of a common sample statistic is distributed.

By engaging students with instruction designed to advance this agenda we simultaneously created a context for pursuing our research goal: to explore and gain insight into students' thinking with respect to making informal inferences on the basis of distributions of values of a sample statistic. We say more about this approach and our particular orientation to characterizing students' thinking when we describe our method of inquiry. But first we elaborate our rationale for the intended learning and instructional goals by situating them within a discussion of the logic of statistical inference.

The logic of statistical inference involves making claims about an underlying population on the basis of the outcome (i.e., a value of the sample statistic of interest) of a single sample that is randomly drawn from that population. This is the logic underlying point estimation in formal inference. We deliberately refer to this here as the *apparent* logic of statistical inference in an effort to highlight our claim that, for statistically naïve students, this logic masks important aspects of inference. The aspects it masks have to do with the logic of interval estimation in formal inference, which allows statisticians to make an estimate about a population parameter with some level of confidence. This latter logic has the idea of sampling distributions at its core; it is the anticipation of the distributional structure of a collection of values of a common sample statistic that dictates how confident we can be that any individual sample (and hence the value of its statistic) is likely to adequately represent the sampled population.

In the larger study from which this report is drawn, our instructional agenda included having students understand this deeper logic of inference. However, we wished to engage students only with informal inference and thus had them work with small collections of values of a sample statistic (i.e., empirical sampling distributions) whose accumulation they could track and which might be manageable for them to structure and construe as distributions.

Participants

Eight liberal-arts-bound students—one 10th-grader, three 11th-graders, and four 12th-graders—enrolled in an introductory statistics course at a suburban high school in the Southeastern United States participated in a 17-session classroom teaching experiment conducted during their fall semester. The high school was located in an upper middle class

suburb. The student population was fairly homogeneous in its racial and socio-economic make-up, consisting largely of white English-speaking adolescents from upper middle class backgrounds.

All students had completed a standard Algebra II course that included a short unit on statistics and probability. This was their only known prior formal instruction in statistics. A written pre-assessment designed to query students' knowledge of sampling and related ideas revealed that students had largely intuitive understandings of the notions of sample and population, consistent with those documented by Watson and Moritz (2000) in a study of students' conceptions of sampling. For instance, there was evidence that many students understood a sample to be a "little part of something".

Method of Inquiry

The study is the second in a sequence of two investigations that employed classroom teaching experiments (Steffe & Thompson, 2000) as the method of inquiry. A centerpiece of the teaching experiment methodology is to use student engagement with instructional tasks as a vehicle for studying their ways of knowing and learning with respect to the concepts targeted in instruction. This method is often used by researchers to generate models of student thinking at a grain size that can usefully posit students' constructive processes and mechanisms that might underlie their conceptual advances. The overarching aim of our research program is to produce cognitive models of the ideas of sampling, variability, and sampling distribution that describe ways of thinking about them that are schematic, imagistic, and dynamic, and that posit hypotheses about their development in relation to students' engagement in classroom instruction (von Glasersfeld, 1995; Thompson & Saldanha, 2000).

Students' understandings and emerging conceptions were investigated in three ways in our study: 1) by tracing their participation in classroom discussions (all instruction was videotaped); 2) by examining their written work; 3) by conducting individual clinical interviews. Three research team members were present in the classroom during all lessons: one author designed and conducted the instruction; the other observed the instructional sessions and generated field notes; a third member operated the video camera(s).

Aspects of Instruction and the Class

Two overarching and related themes permeated instruction throughout the experiment's various phases: 1) the process of randomly selecting samples from a population can be repeated under similar conditions, and 2) assessing a sample's representativeness and inferring a population parameter's value can be based on relative frequency patterns that emerge in collections of values of the sample statistic calculated for similar samples[1]. Instructional

[1] Similar samples share a common size, selection method, and parent population

activities were anchored around these themes, which were intended to support students' developing a distributional perspective of sampling and likelihood (Kahneman & Tversky, 1982). With these themes in mind to guide the experiment's progress, activities and lessons were revised daily according to what the research team perceived as important issues that arose for students in each instructional session. Indeed, the instructional methodology was flexible and responsive to local interactions, allowing for extemporaneous and in situ diversions from the planned instructional activities whenever the instructor deemed that such would advance the overarching instructional agenda. This feature was made possible by the fact that the designer and instructor were one and the same person.

The instructional activities were designed and conducted as discussion-based inquiry-oriented investigations. In accordance with our research and instructional agenda, the mathematical content of the teaching experiment was light on calculations and symbol use, but heavy on explication, description, and connection of ideas[2]. This agenda was enacted by the team members in their on-going interactions with students; we moved to negotiate a culture of sense-making in the classroom by placing a high premium on and promoting pro-active participation such as listening, reflecting, questioning, conjecturing, and explaining and describing one's own and others' thinking about mathematical ideas under discussion.

RESULTS AND DISCUSSIONS

Our report focuses on events that transpired over three consecutive classroom lessons in the initial phase of the teaching experiment. We present our results in two parts: in Part 1 we view the class as a unit of analysis and highlight broad developments in students' participation that point to their thinking as they engaged in a sequence of instructional activities that unfolded in this phase. The activities were designed specifically to support students' abilities to conceive of a collection of values of a sample statistic as a sampling distribution. Given the age of the students and our research agenda, as elaborated above, we were not dealing with theoretical sampling distributions. Instead, students were encouraged to explore empirical sampling distributions made of small collections of values of a given statistic. Nevertheless, we will habitually refer to these as "sampling distribution" in what follows, acknowledging that our use of this term is somewhat an abuse of terminology.

Our analysis in this part characterizes key attentional and discursive shifts that occurred in the classroom and the co-evolution of students' ideas and engagement with instruction. In Part 2 we present a qualitative analysis of individual students' written work on a related set of assessment tasks administered at the conclusion of instruction, supplemented with data from relevant classroom discussions. Our analysis characterizes students' underlying imagery and conceptions in relation to their engagement in the classroom activities and the understandings we moved to foster within them.

[2] Proportions were the most sophisticated calculations used in the course.

Part 1: The Unfolding of Instruction and Students' Engagement

We begin with an overview of the unfolding of instruction. The instructional sequence began with a concrete sampling activity that focused students' attention on the variability among outcomes of samples drawn randomly from populations having a parameter of interest, the value of which was unknown to them. Students then used computer simulations of random sampling to generate collections of values of a sample statistic, and they employed a rule for deciding whether two such collections were similar or dissimilar. The sequence concluded by having students compare collections of values of a sample statistic with the aim of deciding whether a given collection had an unusual distribution, in the sense that collections similar to it would not be expected with high frequency. Because the unfolding of this sequence was tightly coupled with students' engagement, we present the two in tandem as a result.

In the initial sampling activity groups of two or three students hand-drew a small number of random samples from one of three different dichotomous populations of objects, the compositions of which were unknown to them[3]. Students recorded the value of the statistic of interest for each sample (i.e., the number of a particular kind of object in each sample) and began to look for patterns across those values. Discussions centered on what those patterns might suggest about the proportion of objects in the sampled populations, what individual values of the sample statistic might be if the experiment were repeated, and how to decide whether two sets of values of the statistic were similar or dissimilar. In these discussions we first focused students' attention on the variability among values of the sample statistic, and there soon emerged a consensus among students that this variability makes problematic any claim about a population's composition that is based on an individual sample's outcome. This led to the idea of looking at collections of values of the sample statistic, instead. Each group of students drew 10 small samples of equal size from a relatively large population of objects. The groups recorded the corresponding 10 values of their sample statistic of interest and the class investigated how each of these collections of values, as a whole, was distributed.

Moving toward making an inference on the basis of a collection of values

At this point students' discourse began shifting away from describing individual sample outcomes, in terms of the value of the statistic of interest for an individual sample, and toward describing a collection of values of the statistic generated by repeating the sampling process and calculating the statistic's value for each sample. The ensuing discussions focused on how to look at a collection of values of the sample statistic in order to infer the value of the corresponding population parameter of interest. In this case, the parameter was the proportion of one type of object in the underlying dichotomous population.

One pair of students had selected 10 samples of 5 candies each from an opaque sack of red and white candies, the proportion of which was unknown to everyone. After some discussion

[3] This activity was immediately preceded by a whole-class discussion centered on a simulated sampling demonstration in which we broached the ideas of random sample, population, and making an inference from the former to the latter.

of their results and noticing that they had many samples containing more whites than reds (see Figure 1.), students concluded that there were more white candies than red ones in the sack. They based this conclusion on their observation that 80% of the samples were "heavy on the white"—meaning that they contained three or more white candies. We henceforth refer to this collection of values of the sample statistic (see Figure 1) as Result 1. The top row of the table displays the observed values of the sample statistic, whereas each corresponding cell in the bottom row displays the frequency with which that value occurred.

Number of red candies	0	1	2	3	4	5
Number of samples	0	5	3	1	0	1

Figure 1. Result 1: Drawing ten samples of 5 candies each from a population containing 50% red and 50% white candies.

Structuring a collection of a sample statistic's values as a distribution

Thus, students seemed to reason that if a majority of the samples that generated the collection of values each contained a majority of white candies, then this was sufficiently strong evidence to infer that the population also contained a majority of white candies.This line of reasoning suggests that these students were able to coordinate two levels of thinking: one level involves individual samples and their composition; another level involves partitioning the collection of sample compositions in order to ascertain what proportion of them are composed mostly of white candies. This was the first instance in which students displayed an ability to operationalize the criterion for deciding what to infer about the sampled population. It is worth noting that at this point, their inference seemed to be quasi-quantitative—it still lacked a precise measure. They claimed only that the population, like a majority of the samples, contained a majority of white candies.

When the population proportion was subsequently revealed to students, they were surprised to learn that there were actually equal numbers of red and white candies in the sack. In an effort to capitalize on students' apparent surprise and to draw their attention to the issue of how confident one might be about making an inference on the basis of a collection of values of a sample statistic, the instructor raised two key questions at this juncture. The first question referred to the collection of values comprising Result 1, in light of the recently divulged fact that the sampled population had an equal distribution of red and white candies; he asked "might these results be unusual?" He then posed the meta-question "how might we investigate this question?" These questions spawned a series of developments that pushed students to engage with an added layer of conceptual complexity beyond the usual framework of statistical inference. This added layer involved repeating the entire process of selecting multiple samples and recording the value of the statistic for each one (instead of just repeating the selection of a single sample and

recording its statistic's value). This also involved having students think what it means for an entire collection of values of a sample statistic to have an unusual distribution (instead of just thinking what it means for the value of a statistic from a single sample to be unusual). These developments are described in the sections that follow.

Repeating a sampling experiment as a tool for exploring a hypothesis

Following the two questions posed by the instructor, a sustained discussion then ensued about them in which one student eventually suggested repeating multiple times the entire process of selecting 10 samples of 5 candies and recording the 10 values of the sample statistic. Her idea was to compare the resulting collection of 10 values with Result 1: "test it over and over", she said, but without specifying the nature of the proposed comparison.

In order to follow this student's suggestion, at this juncture we introduced the sampling simulation software Prob Sim (Konold & Miller, 1996) as a tool to efficiently simulate repeating the entire experiment of drawing 10 samples of 5 candies from a large population composed of 50% red and 50% white candies. Each time that we simulated collecting 10 samples, we had the program display a frequency distribution of the number of samples having various numbers of red and white candies (see Figure 2) and we led a discussion about these distributions.

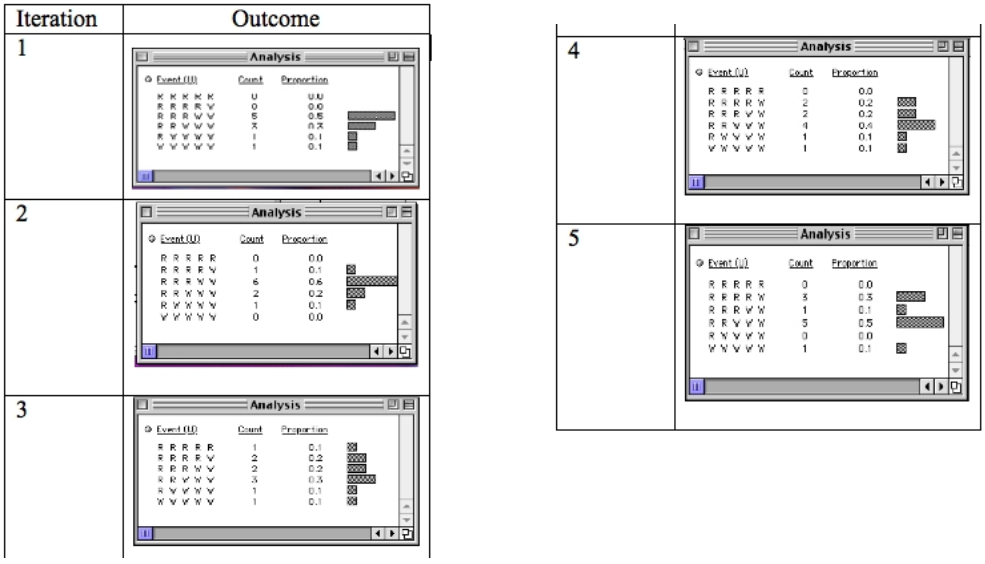


Figure 2. The sequence of outcomes of the simulated candy-sampling experiments.

Discussion of the results of each simulated experiment focused on making sense of and interpreting the information displayed in the Analysis window of the program. Each of the 10 sampling outcomes displayed in this window implies a corresponding collection of 10 values of the sample statistic.

The emergence of a decision rule

For each iteration of the simulated experiment (selecting 10 sample of 5 candies each) the discussion concluded with a class vote taken to decide whether that collection of values of the sample statistic was similar or dissimilar to Result 1 (Figure 1). At first, students were uncertain about how to decide; some expressed opinions that they were unable to justify. We suggested that they think of a collection of these values in terms of the relative weight of parts of its histogram that was displayed in the software window. Through these discussions the characterization "heavy toward the white" emerged as a suitable criterion for deciding whether a collection of values of the sample statistic resembled the one comprising Result 1. This characterization emerged as an informal description of the distribution of Result 1: it is "heavy toward the white" because a majority of the samples contained a majority of white candies. The class applied this criterion over 5 simulations (see Figure 2), and the decisions were recorded and displayed in a table as shown in Figure 3. In the end students agreed that empirical sampling distributions like Result 1 are not unusual once they saw that three of the five simulations produced a distribution of values "similar" to it.

Distribution	1	2	3	4	5
Similar ?	No	No	Yes	Yes	Yes

Figure 3. Record of decisions of whether the simulated process of selecting 10 samples of 5 candies produced a distribution of outcomes similar to Result 1.

Discussion of Part 1

We find it useful to consider the instructional episodes described here as a sequence of phases unfolding out of cycles of interaction between instruction and student engagement and thinking. From our perspective, a first critical development occurred with students' realization—provoked by an orienting cue from instruction—that the variability among values of the sample statistic necessitated a consideration of how collections of such values were distributed in order to infer the underlying population's composition. This led to a second phase of engagement marked by a focus on outcomes of multiple samples and on repeated sampling from each population to accumulate a collection of values of the sample statistic.

In the case of the samples drawn from the population of candies, students seemed to reason about the corresponding collection of values of the sample statistic as a distribution of values; their inferences were based on the relative number of samples in the collection having a majority of white candies. We hypothesize that this reasoning entails thinking of a collection of values of a sample statistic as having a two-tiered structure: on a first level one focuses on individual sample compositions, one quantifies those compositions and develops a sense of the accumulation of the resultant values of the sample statistic; on a second level, one objectifies the

entire collection of these values and partitions it to determine a part's weight relative to the entire collection. This entails quantifying two different attributes—the composition of a collection of values of the sample statistic (at the second level) and the composition of individual samples (at the first level)—and coordinating these quantities so as to not confound them. These conceptual operations and their coordination can be taken to characterize critical aspects of imagining a collection of values of a sample statistic as comprising a distribution. In retrospect, this line of reasoning was crucial to students' continued productive engagement in the instructional activities. We propose that it underlay their eventual ability to agree on a method for comparing entire distributions of values of the sample statistic and to use this method as the basis of a rule for deciding when two such distributions are similar.

These developments, which marked the emergence of a third phase of the episode, were driven by interactions between two critical events: 1) the tension that students experienced when perceiving a discrepancy between their inference and the actual population proportion, and 2) instruction that capitalized on this tension by promoting a culture of inquiry around its resolution (e.g., by provoking students to investigate whether their surprise was warranted). In this third phase students employed the earlier idea of repeatedly sampling and amassing a collection of values of the sample statistic and they re-applied the idea to entire distributions of such values. That is, they were able to repeat the process of structuring a collection of 10 values of the sample statistic (itself the result of repeated sampling) as a distribution and compare it with Result 1—a development facilitated by Prob Sim's presentation format and by the instructor's suggestion to record each collection of values in a frequency table. In this phase of engagement students' focus was thus on entire distributions of 10 values of a sample statistic. As decisions about the similarity between each simulated sampling distribution and Result 1 accumulated, they were recorded and organized (see Figure 3) to facilitate thinking about the collection of those decisions itself as a distribution.

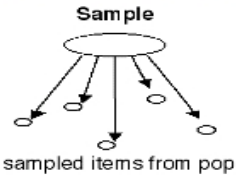
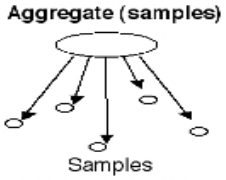
Phase	Activity	Central objects of discourse
1	Select items from population to aggregate a sample Estimate population proportion on the basis of individual sampling outcome (one value of the sample statistic)	Individual samples of 5 items 
2	Select multiple samples to aggregate a collection of values of the sample statistic Estimate population proportion on the basis of the distribution of this collection	Collection of values of the sample statistic 
3	Compare collections of values of the sample statistic Use a decision rule to determine similarity of collections Decide on the relative unusualness of a collection of values of the sample statistic	Collections of values of the sample statistic Collection of similarity decisions

Figure 4. A hypothetical trajectory of the class's development over the unfolding of instruction.

Indeed, students were apparently able to arrive at a collective judgment about the relative unusualness of sampling distributions like Result 1 by considering what proportion of the collection of 5 simulated sampling distributions (each consisting of 10 values of the sample statistic) were "similar" to Result 1. That is, students considered what proportion of the 5 similarity decisions consisted of a "Yes" decision. The reasoning involved in doing so can thus be described as entailing conceptual operations similar to those described above, but applied to a nested object involving an added layer of complexity: a collection of 5 decisions ("Yes" or "No"), each of which itself encapsulates a distribution of values of a sample statistic.

In sum, as students participated in these directed activities and discussions their engagement pointed at a progression in their thinking that can be described in broad terms: they moved from focusing on single sampling outcomes and making an inference on the basis of individual values of the sample statistic, to reasoning about a collection of values of the sample statistic as a distribution of values, and finally to reasoning distributionally about a collection of such collections (of values of the sample statistic). This hypothesized developmental trajectory is summarized in Figure 4.

Part 2: Evidence of Students' Conceptions in an Assessment Activity

Our instructional focus on collections of values of a sample statistic in effect forced

students to organize and structure such collections in ways that enabled them to makes claims about the composition of the underlying populations on the basis of an (empirical) sampling distribution. At the conclusion of the instructional interactions described above students completed a set of task questions designed to query how their engagement with instruction might have shaped their thinking with regard to ideas of sampling and inference.

Figure 5 displays part of the questions that are very similar to those comprising the opening classroom sampling activity. Students were presented with the results of ten samples of five jellybeans each (displayed in the table on the task sheet) that had been selected at random from a large and well-mixed population of red and white jellybeans. Students were not told the composition of the population. Questions 1a and 1b closely parallel those posed in the tangible sampling activity during class; we were interested in the extent to which students were oriented to inferring the population proportion on the basis of individual sampling outcomes and on the basis of the entire collection of sampling outcomes, and we were interested in their reasons for doing either.

Table 1 and Table 2 display the available student responses to Question 1a and Questions 1b, respectively.

Stat assignment 1

Answer these questions on a separate sheet of paper. Copy the question above your answer. Please write complete sentences and explain your thoughts fully.

A large jar is full of red and white jellybeans that are evenly mixed. Ten samples of 5 jellybeans each were selected at random from the jar, the samples had the following outcomes:

1.	red	white	red	white	white
2.	white	white	white	red	white
3.	red	white	white	white	red
4.	white	red	white	white	white
5.	red	red	white	red	red
6.	white	white	white	red	white
7.	red	white	white	white	white
8.	red	red	red	white	white
9.	red	white	white	white	red
10.	white	white	white	white	red

1. Examine how these samples are distributed with regard to the number of red jellybeans in them.

   a. Does any individual sample lead you to believe anything about what fraction of the jar's jellybeans are red? Please explain (i.e., if so, say what and why. Otherwise, say why not).

   b. Does the distribution of the ten samples lead you to believe anything about what fraction of the jar's jellybeans are red? Please explain your answer.

2. Tomorrow you will randomly draw ten samples from this same jar of jellybeans and record the color of each jellybean as you draw it. However, your samples will contain 8 jellybeans instead of 5. Make a list of 10 samples you can reasonably expect to draw. (Use "R" for "red" and "W" for "white".)

Figure 5. Task questions administered at the conclusion of instruction.

Table 1. Students' written responses to Question 1a.

Student	Response
Nicole	No, because one individual sample does not prove a lot.
Sue	No, it doesn't because individual sample might be a rare sample; if so, a fraction of the red jellybeans will be different from the actual fraction.
Kit	Yes, if you just look at any one sample (example sample 4), it would lead you to believe there is about a 1:4 ratio (red: white).
Sarah	Yes, I would normally say the jellybeans are about 3/5 white. Most of the time there are either 3 or 4 white jellybeans out of 5.
Peter	No, not any sample leads you to believe anything about the fraction of red jellybeans. By just looking at one sample, can one be led to believe anything about the jellybeans? It could just be an unusual sample.
Lesley	Most samples, 8 out of 10, had fewer white than red jellybeans. Only 2 out of 10 times would you think there were more red than white jellybeans in the jar.

Table 2. Students' written responses to Question 1b.

Student	Response
Nicole	I'd say no more than 3/5 of the jellybeans are red because only one of the samples have more than that.
Sue	Yes it does. Count out the number of red jellybeans in each sample, and make a list which indicates how many samples in each 0 through 5 possible number of jellybeans. The list shows us which number will occur more likely. Then we can see a fraction.
Kit	It leads you to believe that the ratio would be less reds to more whites.
Sarah	Yes. Most of the time there are 2 or 1 red jellybeans out of five, meaning either 2/5 or 1/5 are red.
Peter	Yes it does. It leads me to believe that the majority of the bag is white and not red. In all samples except one a majority of white beans were taken.
Lesley	One might think that there are fewer reds than whites. Only 2 out of 10 (20%) of the sample have more red than white. Only 36% of the total samples.

Results for Question 1a: Inferences based on a single sample

Only one student (Kit) responded "yes" to Question 1a and employed the single-outcome inferential line of reasoning raised in the discussion surrounding the opening activity of the experiment. Three students—Sarah, Sue, and Peter—answered the question in the negative, their responses suggesting that they thought individual outcomes do not provide enough information about the sampled population to make a trustworthy inference. Two of those students, Sue and Peter, appealed explicitly to the idea that an individual outcome could be unusual. Sue explicitly mentioned the possibility of making an erroneous or unreliable inference as a consequence.

Sarah's and Lesley's responses indicate that they were focused on the collection of outcomes as a whole: both referred to what happened in *most* of the samples as a basis for claiming what might be true of the population, even though Question 1a asked about individual sample outcomes. Sarah ventured to give a specific numerical estimate of the population proportion, thus making a quantitative inference. Lesley fell just short of making such a precise inference, but her response suggests that she thought it obvious that the population contained a majority of red candies.

The difference in their degrees of elaboration notwithstanding, all but Kit's response to Question 1a are consistent with students' thinking that a collection of values of the sample statistic provide a useful, if not preferable, basis for making an inference about the underlying population parameter.

Results for Question 1b: Inferences based on multiple samples

The responses to Question 1b are consistent with those for Question 1a, supporting our claim that students were sensitized to the usefulness of making an inference to the population on the basis of an entire collection of values of the sample statistic. All but one student (Sue) concluded that the population is comprised of fewer red than white jellybeans. Moreover, Sarah's, Peter's and Lesley's responses to both questions, as well as Nicole's response to Question 1b, provide telling evidence of how those students seem to have structured the collection of 10 sampling outcomes to arrive at a conclusion about the population. Their responses suggest the following common features in their thinking: the inference to the population is based on their having quantified each outcome, accumulating values of the sample statistic (i.e., the number of red jellybeans), and partitioning the entire collection of these ten values into some number (or proportion) of them representing the number of samples containing a number (or proportion) of red or white jellybeans.

Images and operations entailed in construing a distribution

We consider a semantic analysis (Vygotsky, 1986) of Peter's response as a paradigmatic example of this line of reason: Peter's conclusion that the population contains more white than red jellybeans ("*the majority of the bag is white and not red*") is based on his having noted that nine of the ten samples each contain a majority of white jellybeans ("*In all samples except one a majority of white beans were taken*"). The explanations provided by the other three students are consistent with this line of reasoning, suggesting a similar structuring of the collection of sampling results. The structuring suggests that these students were able to coordinate two levels of imagery and thinking: one level involves thinking of individual samples and quantifying their composition ("*a majority of white beans were taken*"); another level involves partitioning the collection of accumulated values of the sample statistic in order to ascertain what proportion of

the samples are composed mostly of white candies ("*In all samples except one*"). Thus, at a first level of imagery one focuses on individual samples' compositions and develops a sense of the accumulation of the values of the sample statistic into a collection of them; at a second level the entire collection of values of the sample statistic is objectified and partitioned to determine a part's weight relative to the entire collection. Figure 6 highlights these two levels in regard to Peter's response.

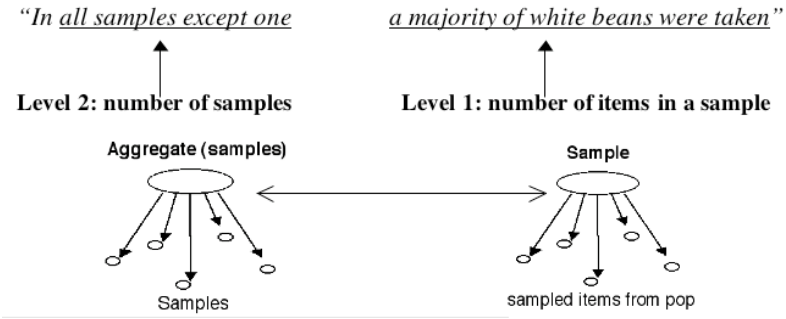


Figure 6. An analysis of Peter's response to Question 1b in terms of two levels of imagery.

This line of reasoning thus entails quantifying two subtly different attributes—the composition of a collection of values of the sample statistic (at the second level), and the composition of individual samples (at the first level)—and coordinating these quantities so as to not confound them. The conceptual operations described above and their coordination can be taken to characterize critical aspects of construing a collection of values of a sample statistic as a distribution.

The following data excerpt is from a classroom discussion centering on the collection of sampling results from this task, after the instructor had organized them into a frequency table on the board (see Figure 7), just like those tables used pervasively by the students in the preceding classroom activities. In the excerpt Peter is explaining to Lesley how to interpret the displayed results so as to make an inference to the sampled population, thus further evidencing his structuring of the collection of sampling results.

# of whites	0	1	2	3	4	5
# of samples	0	1	1	3	5	0

Figure 7. The sampling data organized in a frequency table.

Lesley: "Ok, but I don't understand how you can tell it just from that little thing, right there" (points at table in Figure 7). Ok, 'cause, like…"

Peter: "Alright, see there are 10 samples".

Lesley: "Right".

Peter: "And three—alright, from three, four and five (points at table in Figure 7), if you have three, four, or five whites in … like if you have three whites in one sample that means you have a majority of whites, right?"

Lesley: "Right".

Peter: "Right? And three, four, and five (points at table in Figure 7 and motions with hand from left to right as though counting), so anything on the right side of the table (sweeps hand to right while pointing to table in Figure 7) means you would have a majority of whites. Like…"

Instructor:"So, in these three samples" (points at data value "3"in the second row of the table in Figure 7).

Nicole (to Lesley): "Like, eight of the samples".

Peter (to Lesley): "The numbers under three, four, and five".

David (to Lesley): "Eight of the samples have more than three whites".

Lesley: "Ok".

…

Peter: "It's just eight out of ten, that's 80%".

Peter's explanation of how to consider the tabulated data strongly suggests his having structured the data in terms of a partition: the number or proportion of samples containing a majority of white candies. His imagery of this partition seemed to entail a vivid graphic component, as he noted, and motioned to in a sweeping gesture, that the entries in the right half of the table constituted 80% of the samples. Figure 8 aims to capture Peter's envisioned structuring, which seemed to entail coordinating the two levels of imagery highlighted in Figure 6.

Figure 8. Peter's partitioning of the tabulated sampling outcomes.

Results for Question 2: Anticipating the outcome of multiple samplings

The students' responses to Question 2 are displayed in Table 3. The task in this question was to create a list of ten samples of eight candies each that one would expect to draw from the same population of red and white candies. Our intent was to engage students in a sampling thought-experiment in which they might imagine keeping track of individual outcomes and their accumulation into a collection of outcomes. We were interested in students' abilities to anticipate a reasonable collection of sampling outcomes drawn from the same population, the bases for their predicted outcomes, and whether they were oriented to making connections between their anticipated outcomes and the population inference they made in Question 1.

Table 3. Students' written responses to Question 2.

Sample	Nicole *	Sue *	Kit *	Sarah *	Peter	Lesley	
						Red	White
1	RRRWWWWW	RRWWRWWW	RWWWWWRW	WWWWWRRR	WWWWWRRR	5	3
2	RRWWWWWW	WWWWRWRR	RWRWRWRW	WRWWRWWW	WWWWWWRR	5	3
3	RRRWWWWW	WRWWRRWR	WRRRRRRW	RWWWRWWR	WWWWRRRR	5	3
4	RRWWWWWW	RWWWWWWR	WWWWRWRW	RRRWWWWW	WWWWRRRR	2	6
5	RRRRRWWW	WWRRWRWR	RWWRWRWW	RWWWWWWW	WWWWWRRR	2	6
6	RRWWWWWW	WWWRWRWR	WWWRWWRR	WWRRWWWW	WWWRRRRR	2	6
7	RRWWWWWW	WRWWRWWR	RRWWWRWW	WWRWRWWR	WWRRRRRR	3	5
8	RRRRWWWW	RWWRRRRW	RWWWRRRW	RWWWWWWR	WWWWWWRR	5	3
9	RRRWWWWW	WRWWRWWW	WRWWWRWW	RWWRWRWW	WWWWRRRR	4	4
10	RRWWWWWW	RWRWWRWW	WRRWWRWW	WWWRRRRW	WWWWWRRR	4	4
						total R	total W
						37	43
						46%	54%

(*) The student's list is weighted toward samples having more white than red jellybeans.

Each column of Table 3 lists an individual student's response, and captures both his/her ordering of the sample elements and of the samples themselves. All students except Lesley listed the samples by writing sequences of the symbols "R" or "W" as suggested in the task directive. Lesley's response entailed recording only the number of red and white jellybeans expected in each sample, thus suggesting a focus on quantifying the sample compositions outright. If we consider each student's list as a distribution of values of the sample statistic (number of whites jellybeans), we can see that all but two are definitively weighted toward samples having more white than red beans. The exceptions are Peter's and Lesley's lists, which are both borderline cases; exactly half, rather than a majority, of each of their samples contain a majority of white jellybeans. However, the following excerpt from a classroom discussion in which the instructor provoked students to explain their responses to Question 2 indicates that Peter and Lesley had both tried to construct lists of sampling outcomes that would reflect a sampled population comprised of mostly white jellybeans.

> **Instructor**: "Ok. Now what does this (points at frequency table in Figure 7), uhh what did we say that this suggests, about the jar?"
>
> **Peter**: "A majority of them are white".
>
> **Instructor**: "Majority of them are white. All right. So, uhh if we picked eight and, in fact, there are a majority of white in the jar".
>
> **Peter**: "Like".
>
> **Instructor**: "Then what would you expect over the long run?"

> **Peter**: "What I was doing when I was making my samples was, I was—what I got from this that you gave me (points at the task questionnaire), there's a majority of white. So when I was doing mine, I was, I guess I was trying to make sure that there was, like, a majority of white".
>
> **Instructor**: "Ok".
>
> **Lesley**: "Yeah".

Thus, overall, students' anticipated sampling outcomes were consistent with their inference in Question 1b that the sampled population contained more white than red jellybeans.

Discussion of Part 2

These results, particularly the evidence for students' construal of a collection of values as a distribution, are informative as to how one might productively structure such collections in order to make an inference to the sampled population on the basis of a distribution of sample statistics. These results also indicate that, in this culminating task of the first phase of the experiment, the students' conceptions of a collection of values of a sample statistic seemed well aligned with the conception we targeted and had moved to foster in instruction.

An alternative view of repeated sampling and its aggregation

In addition to these results there is some evidence from this phase to suggest that students could well have developed an alternative way of thinking about the collection of values of the sample statistic as a basis for making an inference to the sampled population. As a case in point, we consider Lesley's responses to Question 1b and Question 2 in their entirety:

> **Lesley**: "One might think that there are fewer reds than whites. Only 2 out of 10 (20%) of the sample have more red than white. Only 36% of the *total* samples."

We already noted how the first part of Lesley's response suggests a way in which she might have structured the given collection of sampling outcomes. But the last sentence of her response: "only 36% of the *total* samples", does not apparently follow from the first part and would seem to suggest a different interpretation of her thinking in this respect. We note the emphasis that Lesley placed on the word "*total*" in this part of the response (underlined by her) and that she used the same term in her response to Question 2, evident in the last column of Table 3. This part of Lesley's response is consistent with her having thought of the given collection of ten samples as constituting one large sample, the composition of which would enable her to make an inference to the underlying population. Indeed, Lesley's complete solution to Question 1b entailed the work displayed in Figure 9: it clearly demonstrates that Lesley also tallied up the number of red and white jellybeans in each of the ten given samples to arrive at a total of 18 red and 32 white jellybeans.

	R	W
1. red white red white white	2	3
2. white white white red white	1	4
3. red white white white red	2	3
4. white red white white white	1	4
5. red red white red red	*4	1
6. white white white red white	1	4
7. red white white white white	1	4
8. red red red white white	*3	2
9. red white white white red	2	3
10. white white white white red	+1	+4
	18	32

(*)Samples containing "more red than white".

Figure 9. Lesley's work on Question 1b, as it appeared on her copy of the task sheet.

Thus, "36% of the total samples" was Lesley's way of expressing the fact that 18 out of a total of 50 jellybeans were red. Lesley employed the same line of reasoning when answering Question 2 (see Table 3), as further evidenced in the following brief excerpt from the classroom discussion around that question.

Instructor: "Ok. Uhh, what did you do to look at all 10 samples to uhh as one collection?" (3 second silence)

Instructor: "Did you just eyeball it? Or did you summarize it somehow?"

Lesley: "Oh, well I added up all the reds and all the whites".

Instructor: "Ok, so you count…"

Lesley: "And looked at the percentage".

Instructor: "So you looked at the numbers of each?"

Lesley: "Right".

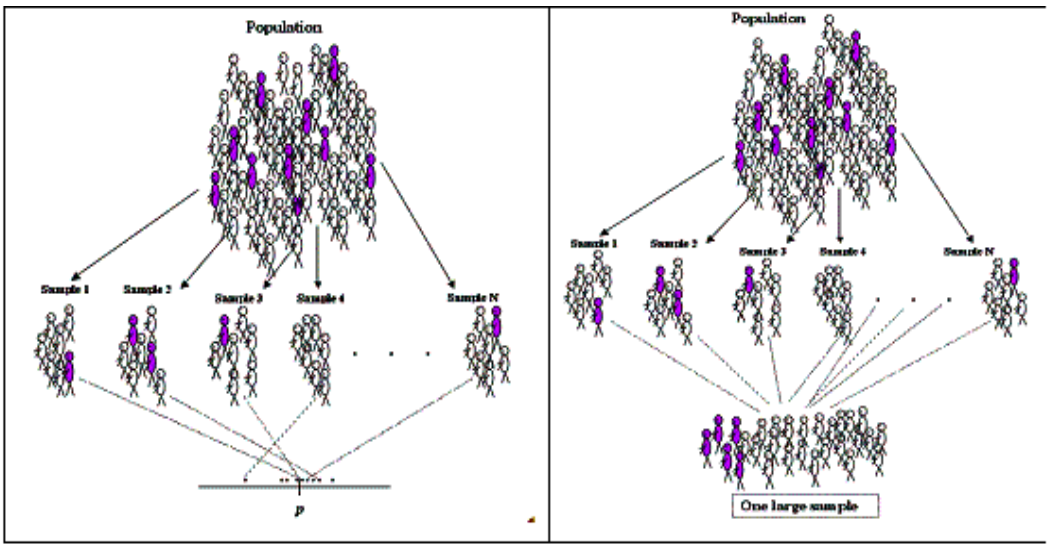


Figure 10. Two views of repeated sampling and its aggregation.

Figure 10 conveys the difference between two ways of thinking about the repeated sampling and its results: the view depicted in the left panel entails an understanding that individual samples are selected, the value of the sample statistic is calculated for each distinct sample, and that these values accumulate along a range of possibilities to form a collection of values that is internally diverse. We view this conception as proto-distributional, in the sense that it can form a useful basis for structuring the collection of values of the statistic so as to construe it as a sampling distribution. Indeed, our instructional agenda aimed to support students in developing a conception of sampling aligned with this one. The right panel, on the other hand, depicts an understanding of repeated sampling as the accumulation of one large sample, for which one might then calculate the value of the statistic.

Lesley's work suggests that she entertained both of these conceptions, and that the distinction between them may have been ambiguous to her. Although Lesley is the only student who displayed evidence of this, to us it nevertheless underscores the instructional challenge of how to support students in conceiving what is being aggregated in a repeated sampling activity and in making inferences on the basis of a distribution of sample statistics.

CONCLUSION

Our research goal was to gain insight into the thinking and conceptual operations that might be entailed in making a statistical inference on the basis of a collection of values of a sample statistic. Our research method involved engaging students with classroom instructional activities, thus creating a context for studying their thinking in relation to engagement with instruction. At the core of our instructional approach was the idea of repeatedly selecting a sample from a population, calculating the value of a common statistic for each sample, and accruing a collection of such values. Our approach deliberately departed from the standard method of generating sampling distributions via the appropriate theoretical probability models. Instead we had students generate small collections of a statistic's values and we employed simulations as a didactic tool. We aimed for students to think of these values as each representing a distinct sample and to think of the resulting collection as representing an aggregate of distinct sampling outcomes, the whole of which could be organized and structured as a distribution.

The results presented in this report demonstrate that students can learn to make inferences to a population on the basis of conceiving a *collection* of values of a sample statistic as a distribution—an idea that lies at the heart of statistical inference, yet is often only implicit in instruction. Moreover, our study has enabled us to move toward proposing a viable model of the conceptual operations (illustrated in Figure 6) entailed in construing a collection of values of a sample statistic as a distribution. In addition to these results, we saw in one student's—Lesley—work an alternative and very reasonable way of construing the repetitive sampling scheme and the resulting collection that accrues. In our interpretation, Lesley had an image of

multiple samples accruing to form one large sample, the composition of which would form the basis of an inference to the underlying population. Lesley thus appeared to have assimilated the repeated selection of samples to a scheme akin to the usual logic of statistical inference. In our view, the importance of Lesley's work is that it points to an ambiguity that teachers might reasonably expect students to experience when attempting to engage them with instructional activities of the type described here. Indeed, the tension that Lesley may have experience between two competing views of repeated sampling and its aggregation hints at an implication for teaching; it underscores the instructional challenge of helping students tease apart two subtly different, yet important, understandings of repeated sampling, only one of which supports reasoning about sampling distributions. In teaching students to understand inference as embedded within the idea of sampling distributions, we would do well not to assume that the purpose of repeated sampling is transparent to students, nor to assume that they will conceive the resulting collection of values that accrues in ways that supporting thinking of them as a distribution of a sample statistics.

We close our discussion by commenting on a limitation of this study that points to future research; our exploration of student's propensities to make population inferences on the basis of empirical sampling distributions was limited to a context involving much instructional scaffolding that unfolded within a relatively short period of time. This necessarily constrains us from making claims about the longevity and sustainability of their demonstrated abilities to make such inferences beyond this temporal period, in subsequent phases of the larger teaching experiment, and outside of such directly supported environments. The robustness and limitations of their abilities outside of these parameters are currently unknown; their exploration will have to form the basis of a future report.

ACKNOWLEDGEMENT

Research reported in this paper was supported by National Science Foundation Grant No. REC-9811879. Any conclusions and recommendations stated here are those of the authors and do not necessarily reflect official positions of NSF. The first author gratefully acknowledges the support provided by the College of Liberal Arts and Sciences of Portland State University for the production of this report.

REFERENCES

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *American Mathematical Monthly*, 104, 801-823.

delMas, R. C., Garfield, J., & Chance, B. L. (1999). Exploring the role of computer simulations in developing understanding of sampling distributions. Paper presented at the *American Educational Research Association Conference*, Montreal.

Glasersfeld, E. v. (1995). *Radical constructivism: A way of knowing and learning*. London: Falmer Press.

Kahneman, D., & Tversky, A. (1982). Variants of uncertainty. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases (*pp. 509-521). New York: Cambridge University Press.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psycholog*, 3, 430-454.

Konold, C. (1989). Informal conceptions of probability. *Cognition and Instruction*, 6(1), 59-98.

Konold, C., & Miller, C. (1996). *Prob Sim*. Computer Program. Amherst, MA.

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education*, 33 (4), 259-289.

NCTM. (2000). *Principles and standards for school mathematics*. Reston, VA: National Council of Teachers of Mathematics.

Piaget, J., & Inhelder, B. (1951). *La genèse de l'idée de hasard chez l'enfant* (The origin of chance in children). Paris: Presses Universitaires de France.

Rubin, A., Bruce, B., & Tenney, Y. (1991). Learning about sampling: Trouble at the core of statistics. In D. Vere-Jones (Ed), *Proceedings of the Third International Conference on Teaching Statistics* (Vol. 1, pp. 314-319). Dunedin, New Zealand: International Statistical Institute.

Saldanha, L. A. (2004). *"Is this sample unusual?": An investigation of students exploring connections between sampling distributions and statistical inference*. Unpublished Doctoral Dissertation.Vanderbilt University.

Saldanha, L. A. & Thompson, P. W. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics*, 51, 257-270.

Schwartz, D. L., Goldman, S. R., Vye, N. J., & Barron, B. J. (1998). Aligning everyday and mathematical reasoning: The case of sampling assumptions. In S. P. Lajoie (Ed.), *Reflections on statistics: learning, teaching, and assessment in grades K-12* (pp. 233-273). Mahwah, NJ: Lawrence Erlbaum.

Sedlmeier, P. (1999). *Improving statistical reasoning: Theoretical models and practical implications*. Mahwah, NJ: Lawrence Erlbaum.

Sedlmeier, P., & Gigerenzer, G. (1997). Intuitions about sample size: The empirical law of large numbers. *Journal of Behavioral Decision Making*, 10, 33-51.

Shaughnessy, J. M., Watson, J., Moritz, J., & Reading, C. (1999). School students' acknowledgment of statistical variation. Paper presented at the *Research Pre- session Symposium of the 77th Annual NCTM Conference*, San Francisco,CA.

Steffe, L. P., & Thompson, P. W. (2000). Teaching experiment methodology: Underlying principles and essential elements. In A. E. Kelly, & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 267-306). Mahwah, NJ: Lawrence Erlbaum.

Thompson, P. W., Saldanha, L. A., & Liu. Y. (2004). Why statistical inference is hard to understand. Paper presented at the *Annual Meeting of the American Educational Research Association*, San Diego, April 2004.

Thompson, P. W., & Saldanha, L. A. (2003). Fractions and multiplicative reasoning. In J. Kilpatrick, G. Martin, & D. Schifter (Eds.), *Research companion to the principles and standards for school mathematics* (pp. 95- 113). Reston, VA: NCTM.

Thompson, P. W., & Saldanha, L. A. (2000). Epistemological analyses of mathematical ideas: A research methodology. In M. L. Fernandez (Ed.), *Proceedings of the Twenty Second Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (Vol. 2, pp. 403-408), Columbus, OH: ERIC Clearinghouse for Science, Mathematics, and Environmental Education.

Vygotsky, L. S. (1986). Thought and language. Cambridge, MA: MIT Press.

Watson, J. M., & Moritz, J. B. (2000). Developing concepts of sampling. *Journal for Research in Mathematics Education*, 31 (1), 44-70.

Well, A. D., Pollatsek, A., & Boyce, S. J. (1990). Understanding the effects of sample size on the variability of the mean. *Journal of Organizational Behavior and Human Decision Processes*, 47, 289-312.

Author : **Luis A. Saldanha**

E-mail : saldanha@pdx.edu

Address : Department of Mathematics and Statistics
Portland State University
P.O. Box 751, Portland, OR 97207, U.S.A.

Phone : (503) 725-8295

Fax : (503) 725-3661

Author : **Patrick W. Thompson**

E-mail : pat.thompson@asu.edu

Address : CRESMET, Arizona State University
P.O. Box 873604, Tempe, AZ 85287, U.S.A.

Phone : (480) 727-8885

Fax : (480) 965-5993